A Variant of AIC based on the Bayesian Marginal Likelihood

Yuki Kawakubo[*] Tatsuya Kubokawa[†] and Muni S. Srivastava[‡]

Abstract

We propose information criteria that measure the prediction risk of a predictive density based on the Bayesian marginal likelihood from a frequentist point of view. We derive criteria for selecting variables in linear regression models, assuming a prior distribution of the regression coefficients. Then, we discuss the relationship between the proposed criteria and related criteria. There are three advantages of our method. First, this is a compromise between the frequentist and Bayesian standpoints because it evaluates the frequentist's risk of the Bayesian model. Thus, it is less influenced by a prior misspecification. Second, the criteria exhibits consistency when selecting the true model. Third, when a uniform prior is assumed for the regression coefficients, the resulting criterion is equivalent to the residual information criterion (RIC) of Shi and Tsai (2002).

Keywords: AIC; BIC; Consistency; Kullback–Leibler divergence; Linear regression model; Residual information criterion; Variable selection.

1 Introduction

The problem of selecting appropriate models has been studied extensively in the literature since the work of Akaike (1973, 1974), who derived the so-called Akaike information criterion (AIC). There are several approaches to solving the model selection problem: information criteria, such as the AIC or BIC (Schwarz, 1978); shrinkage methods, such as the lasso (Tibshirani, 1996); Bayesian techniques; among others. With regard to Bayesian techniques, O'Hara and Sillanpaa (2009) provide a good review of the key works in this field, including Kuo and Mallick (1998), Dellaportas et al. (1997), and George and McCulloch (1993, 1997). In addition, a Bayesian lasso procedure based on a spike-and-slab prior has attracted much

[*]Faculty of Law, Politics, and Economics, Chiba University, 1-33, Yayoi-cho, Inage-ku, Chiba, 263-8522, JAPAN, (E-mail: `kawakubo@chiba-u.jp`)

[†]Faculty of Economics, University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, JAPAN, (E-mail: `tatsuya@e.u-tokyo.ac.jp`)

[‡]Department of Statistics, University of Toronto, 100 St George Street, Toronto, Ontario, CANADA M5S 3G3, (E-mail: `srivasta@utstat.toronto.edu`)

recent attention (Xu and Ghosh, 2015). However, although these methods are useful and important, we focus on information criteria in this study.

Two of the fundamental information criteria are the AIC and the BIC. Because the AIC and its variants are based on the risks of predictive densities with respect to the Kullback–Leibler (KL) divergence, they are able to select good models in terms of their predictive ability. In fact, Shibata (1981), Shao (1997), and others have shown that the model selected by the AIC minimizes the prediction error asymptotically. However, it is known that AIC-type criteria do not have the property of consistency; that is, the probability that the criteria will select the true model does not converge to 1. On the other hand, the BIC, based on the Bayesian marginal likelihood, does exhibit consistency in certain specific models (Nishii, 1984), but does not select models as effectively in terms of their predictive ability. Therefore, we propose a hybrid of the AIC and the BIC that uses the empirical Bayesian method, which has the property of consistency, but also selects models well in terms of predictive ability.

Our approach is to measure the prediction risk of a predictive density based on the Bayesian marginal likelihood from a frequentist point of view. Specifically, we focus on the variable selection problem for normal linear regression models, assuming a prior distribution of the regression coefficients in order to derive the criterion. In Section 3, we consider two prior distributions, namely the normal distribution and the uniform prior distribution. There are three advantages of our method. First, this is a compromise between the frequentist and Bayesian standpoints because it evaluates the frequentist's risk of the Bayesian model. Thus, the method should be less influenced by a prior misspecification. Second, the criteria exhibits consistency when selecting the true model. At the same time, our proposed criteria can select a good model, in the sense that the prediction risk is small, because the criteria are based on the KL divergence. Lastly, a non-informative improper prior can be also used to construct criteria using our approach. If we consider the Bayesian risk, which is the KL risk, integrated out with respect to the parameters based on the prior distribution, it diverges when the prior is improper. On the other hand, when we assume a uniform improper prior for the regression coefficients in the normal linear regression model, we can formally derive the marginal likelihood. In this case, the resulting marginal likelihood is the so-called residual likelihood (Patterson and Thompson, 1971), and the proposed information criterion is equivalent to the residual information criterion (RIC) of Shi and Tsai (2002). Thus, our approach can be considered a theoretical justification of the RIC.

The rest of the paper is organized as follows. In Section 2, we provide a unified framework, which we use to derive the proposed information criteria, that can produce various information criteria, including the AIC, BIC, and RIC. Then, we propose a new approach that uses the Bayesian marginal likelihood in the general framework, and compare various information criteria that are based on a Bayesian model. In Section 3, we derive our information criteria for the variable selection problem in a normal linear regression model, assuming a prior distribution of the regression coefficients. In this section, we also prove the consistency of the criteria. In Section 4, we use simulations to verify the numerical performance of the proposed criteria. Lastly, Section 5 concludes the paper.

2

2 General

2.1 Conventional information criteria

In this section, we describe the concept of information criteria from a general point of view. Let $\boldsymbol{y}$ be an n-variate observable random vector, with density $f(\boldsymbol{y}|\boldsymbol{\omega})$ for a vector of unknown parameters $\boldsymbol{\omega}$. Let $\hat{f}(\widetilde{\boldsymbol{y}};\boldsymbol{y})$ be a predictive density of $f(\widetilde{\boldsymbol{y}}|\boldsymbol{\omega})$, where $\widetilde{\boldsymbol{y}}$ is an n-variate independent replication of $\boldsymbol{y}$. Here, we evaluate the predictive performance of $\hat{f}(\widetilde{\boldsymbol{y}};\boldsymbol{y})$ in terms of the following risk:

$$R(\boldsymbol{\omega};\hat{f}) = \int \left[\int \log \left\{ \frac{f(\widetilde{\boldsymbol{y}}|\boldsymbol{\omega})}{\hat{f}(\widetilde{\boldsymbol{y}};\boldsymbol{y})} \right\} f(\widetilde{\boldsymbol{y}}|\boldsymbol{\omega})\mathrm{d}\widetilde{\boldsymbol{y}} \right] f(\boldsymbol{y}|\boldsymbol{\omega})\mathrm{d}\boldsymbol{y}. \tag{1}$$

Because this is interpreted as a risk with respect to the KL divergence, we call it the KL risk. The spirit of AIC suggests that we can provide an information criterion for model selection as an (asymptotically) unbiased estimator of the information, as follows:

$$\begin{aligned} I(\boldsymbol{\omega};\hat{f}) &= \iint -2\log\{\hat{f}(\widetilde{\boldsymbol{y}};\boldsymbol{y})\}f(\widetilde{\boldsymbol{y}}|\boldsymbol{\omega})f(\boldsymbol{y}|\boldsymbol{\omega})\mathrm{d}\widetilde{\boldsymbol{y}}\mathrm{d}\boldsymbol{y} \\ &= E_{\boldsymbol{\omega}}\left[-2\log\{\hat{f}(\widetilde{\boldsymbol{y}};\boldsymbol{y})\} \right], \end{aligned} \tag{2}$$

which is part of (1) (multiplied by 2), where $E_{\boldsymbol{\omega}}$ denotes the expectation, with respect to the distribution, of $f(\widetilde{\boldsymbol{y}},\boldsymbol{y}|\boldsymbol{\omega}) = f(\widetilde{\boldsymbol{y}}|\boldsymbol{\omega})f(\boldsymbol{y}|\boldsymbol{\omega})$. Let $\Delta = I(\boldsymbol{\omega};\hat{f}) - E_{\boldsymbol{\omega}}[-2\log\{\hat{f}(\boldsymbol{y};\boldsymbol{y})\}]$. Then, the AIC variant based on the predictor $\hat{f}(\widetilde{\boldsymbol{y}};\boldsymbol{y})$ is defined by

$$\mathrm{IC}(\hat{f}) = -2\log\{\hat{f}(\boldsymbol{y};\boldsymbol{y})\} + \widehat{\Delta}, \tag{3}$$

where $\widehat{\Delta}$ is an (asymptotically) unbiased estimator of Δ and $\hat{f}(\boldsymbol{y};\boldsymbol{y})$ is the value of the function $\hat{f}(\widetilde{\boldsymbol{y}};\boldsymbol{y})$ evaluated at $\widetilde{\boldsymbol{y}} = \boldsymbol{y}$.

Note that $\mathrm{IC}(\hat{f})$ produces the AIC and BIC for specific predictive densities.

(AIC) Use $\hat{f}(\widetilde{\boldsymbol{y}};\boldsymbol{y}) = f(\widetilde{\boldsymbol{y}}|\widehat{\boldsymbol{\omega}})$ as the maximum likelihood estimator $\widehat{\boldsymbol{\omega}}$ of $\boldsymbol{\omega}$. Then, $\mathrm{IC}(f(\widetilde{\boldsymbol{y}}|\widehat{\boldsymbol{\omega}}))$ is the exact AIC, or is the corrected AIC suggested by Sugiura (1978) and Hurvich and Tsai (1989), which is approximated by the AIC of Akaike (1973, 1974) as $-2\log\{f(\boldsymbol{y}|\widehat{\boldsymbol{\omega}})\} + 2\dim(\boldsymbol{\omega})$.

(BIC) Use $\hat{f}(\widetilde{\boldsymbol{y}};\boldsymbol{y}) = f_{\pi_0}(\widetilde{\boldsymbol{y}}) = \int f(\widetilde{\boldsymbol{y}}|\boldsymbol{\omega})\pi_0(\boldsymbol{\omega})\mathrm{d}\boldsymbol{\omega}$ as a proper prior distribution $\pi_0(\boldsymbol{\omega})$. It is evident that $I(\boldsymbol{\omega};f_{\pi_0}) = E_{\boldsymbol{\omega}}[-2\log\{f_{\pi_0}(\boldsymbol{y})\}]$. Thus, we have $\Delta = 0$, so that $\mathrm{IC}(f_{\pi_0}) = -2\log\{f_{\pi_0}(\boldsymbol{y})\}$, which is the Bayesian marginal likelihood. Note that $-2\log\{f_{\pi_0}(\boldsymbol{y})\}$ is approximated by $\mathrm{BIC} = -2\log\{f(\boldsymbol{y}|\widehat{\boldsymbol{\omega}})\} + \log(n)\cdot\dim(\boldsymbol{\omega})$.

2.2 Proposed approach

The criterion $\mathrm{IC}(\hat{f})$ in (3) can produce not only the conventional AIC and BIC, but also various other criteria. Hereafter, we consider that $\boldsymbol{\omega}$ is divided as $\boldsymbol{\omega} = (\boldsymbol{\beta}^t, \boldsymbol{\theta}^t)^t$, for a p-dimensional parameter vector of interest $\boldsymbol{\beta}$, and a q-dimensional nuisance parameter vector

$\boldsymbol{\theta}$, respectively. We assume that $\boldsymbol{\beta}$ has prior density $\pi(\boldsymbol{\beta}|\boldsymbol{\lambda}, \boldsymbol{\theta})$, with hyperparameter $\boldsymbol{\lambda}$. The model is given as follows:

$$\boldsymbol{y}|\boldsymbol{\beta} \sim f(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{\theta}),$$
$$\boldsymbol{\beta} \sim \pi(\boldsymbol{\beta}|\boldsymbol{\lambda}, \boldsymbol{\theta}),$$

where $\boldsymbol{\theta}$ and $\boldsymbol{\lambda}$ are estimated from the data. An inference based on such a model is called an empirical Bayes procedure. Here, we consider the predictive density $\hat{f}(\widetilde{\boldsymbol{y}}; \boldsymbol{y})$ as

$$\hat{f}(\widetilde{\boldsymbol{y}}; \boldsymbol{y}) = f_\pi(\widetilde{\boldsymbol{y}}|\widehat{\boldsymbol{\lambda}}, \widehat{\boldsymbol{\theta}}) = \int f(\widetilde{\boldsymbol{y}}|\boldsymbol{\beta}, \widehat{\boldsymbol{\theta}})\pi(\boldsymbol{\beta}|\widehat{\boldsymbol{\lambda}}, \widehat{\boldsymbol{\theta}})\mathrm{d}\boldsymbol{\beta}$$

for some estimators, $\widehat{\boldsymbol{\lambda}}$ and $\widehat{\boldsymbol{\theta}}$. Then, the information given in (2) is

$$I(\boldsymbol{\omega}; f_\pi) = \iint -2\log\{f_\pi(\widetilde{\boldsymbol{y}}|\widehat{\boldsymbol{\lambda}}, \widehat{\boldsymbol{\theta}})\}f(\widetilde{\boldsymbol{y}}|\boldsymbol{\beta}, \boldsymbol{\theta})f(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{\theta})\mathrm{d}\widetilde{\boldsymbol{y}}\mathrm{d}\boldsymbol{y}, \tag{4}$$

and the resulting information criterion is

$$\mathrm{IC}(f_\pi) = -2\log\{f_\pi(\boldsymbol{y}|\widehat{\boldsymbol{\lambda}}, \widehat{\boldsymbol{\theta}})\} + \widehat{\Delta}_{f_\pi}, \tag{5}$$

where $\widehat{\Delta}_{f_\pi}$ is an (asymptotically) unbiased estimator of $\Delta_{f_\pi} = I(\boldsymbol{\omega}; f_\pi) - E_{\boldsymbol{\omega}}[-2\log\{f_\pi(\boldsymbol{y}|\widehat{\boldsymbol{\lambda}}, \widehat{\boldsymbol{\theta}})\}]$.

There are three motivations for considering the information $I(\boldsymbol{\omega}; f_\pi)$ in (4) and the information criterion $\mathrm{IC}(f_\pi)$ in (5).

First, the precision of the Bayesian predictor $f_\pi(\widetilde{\boldsymbol{y}}|\widehat{\boldsymbol{\lambda}}, \widehat{\boldsymbol{\theta}})$ is characterized by the risk $R(\boldsymbol{\omega}; f_\pi)$ in (1), which is based on a frequentist point of view. On the other hand, the Bayesian risk is defined by

$$r(\boldsymbol{\psi}; \hat{f}) = \int R(\boldsymbol{\omega}; \hat{f})\pi(\boldsymbol{\beta}|\boldsymbol{\lambda}, \boldsymbol{\theta})\mathrm{d}\boldsymbol{\beta}, \tag{6}$$

which measures the prediction error of $\hat{f}(\widetilde{\boldsymbol{y}}; \boldsymbol{y})$, under the assumption that the prior information is correct, where $\boldsymbol{\psi} = (\boldsymbol{\lambda}^t, \boldsymbol{\theta}^t)^t$. The resulting Bayesian criteria, such as the PIC (Kitagawa, 1997) or the DIC (Spiegelhalter et al., 2002), are sensitive to a prior misspecification because they depend on the prior information. However, because $R(\boldsymbol{\omega}; f_\pi)$ can measure the prediction error of the Bayesian model from a frequentist standpoint, the resulting criterion $\mathrm{IC}(f_\pi)$ is less influenced by a prior misspecification.

Second, this criterion has the property of consistency. In Section 3, we derive criteria for the variable selection problem in normal linear regression models, and prove that the criteria select the true model with probability tending to one. The BIC and marginal likelihood are known to exhibit consistency, while most AIC-type criteria are not consistent. However, AIC-type criteria can choose a good model in the sense of minimizing the prediction error. Our proposed criterion should include both properties, namely consistency when selecting the parameters of interest $\boldsymbol{\beta}$, and the tendency to select a good model in terms of its predictive ability.

Lastly, we can construct the information criterion $\mathrm{IC}(f_\pi)$ even when the prior distribution of $\boldsymbol{\beta}$ is improper, because the information $I(\boldsymbol{\omega}; f_\pi)$ in (4) can be defined formally for the corresponding improper marginal likelihood. However, because the Bayesian risk $r(\boldsymbol{\psi}; f_\pi)$ does not exist for the improper prior, we cannot obtain the corresponding Bayesian criteria or use the Bayesian risk. Note that our criterion is equivalent to the residual information criterion (RIC) of Shi and Tsai (2002) if we assume a uniform prior on the regression coefficients. In general, the marginal likelihood based on an improper prior depends on an arbitrary scalar constant, which can be included in the prior, but that might be problematic when selecting the model. However, the criterion based on our approach, using a uniform prior, can work as a variable selection criterion. This is discussed further in Remark 1 in Section 3.

2.3   Other information criteria based on a Bayesian model

To clarify our proposed approach, we first explain related information criteria that are based on a Bayesian model. When the prior distribution $\pi(\boldsymbol{\beta}|\boldsymbol{\lambda}, \boldsymbol{\theta})$ is proper, we can treat the Bayesian prediction risk $r(\boldsymbol{\psi}; \hat{f})$ in (6). When $\boldsymbol{\psi} = (\boldsymbol{\lambda}^t, \boldsymbol{\theta}^t)^t$ is known, the predictive density $\hat{f}(\widetilde{\boldsymbol{y}}; \boldsymbol{y})$ that minimizes $r(\boldsymbol{\psi}; \hat{f})$ is the Bayesian predictive density (posterior predictive density) $f_\pi^*(\widetilde{\boldsymbol{y}}|\boldsymbol{y}, \boldsymbol{\psi})$, given by

$$\int f(\widetilde{\boldsymbol{y}}|\boldsymbol{\beta}, \boldsymbol{\theta})\pi(\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{\lambda}, \boldsymbol{\theta})\mathrm{d}\boldsymbol{\beta} = \frac{\int f(\widetilde{\boldsymbol{y}}|\boldsymbol{\beta}, \boldsymbol{\theta})f(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{\theta})\pi(\boldsymbol{\beta}|\boldsymbol{\lambda}, \boldsymbol{\theta})\mathrm{d}\boldsymbol{\beta}}{\int f(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{\theta})\pi(\boldsymbol{\beta}|\boldsymbol{\lambda}, \boldsymbol{\theta})\mathrm{d}\boldsymbol{\beta}}.$$

When $\boldsymbol{\psi}$ is unknown, we can consider the Bayesian risk of the plug-in predictive density $f_\pi^*(\widetilde{\boldsymbol{y}}|\boldsymbol{y}, \widehat{\boldsymbol{\psi}})$. In this case, the resulting criterion is known as the predictive likelihood (Akaike, 1980a) or the PIC (Kitagawa, 1997). The deviance information criterion (DIC) of Spiegelhalter et al. (2002) and the Bayesian predictive information criterion (BPIC) of Ando (2007) are related criteria based on the Bayesian prediction risk $r(\boldsymbol{\psi}; \hat{f})$.

Akaike's Bayesian information criterion (ABIC) (Akaike, 1980b) is another information criterion based on the Bayesian marginal likelihood, given by

$$\mathrm{ABIC} = -2\log\{f_\pi(\boldsymbol{y}|\widehat{\boldsymbol{\lambda}})\} + 2\dim(\boldsymbol{\lambda}),$$

where the nuisance parameter $\boldsymbol{\theta}$ is not considered. The ABIC measures the following KL risk:

$$\int \left[\int \log\left\{ \frac{f_\pi(\widetilde{\boldsymbol{y}}|\boldsymbol{\lambda})}{f_\pi(\widetilde{\boldsymbol{y}}|\widehat{\boldsymbol{\lambda}})} \right\} f_\pi(\widetilde{\boldsymbol{y}}|\boldsymbol{\lambda})\mathrm{d}\widetilde{\boldsymbol{y}} \right] f_\pi(\boldsymbol{y}|\boldsymbol{\lambda})\mathrm{d}\boldsymbol{y},$$

which is not the same as either $R(\boldsymbol{\omega}; \hat{f})$ or $r(\boldsymbol{\psi}; \hat{f})$. The ABIC is used to choose the hyperparameter $\boldsymbol{\lambda}$ in the same manner as the AIC. However, note that the ABIC works as a model selection criterion for $\boldsymbol{\beta}$ because it is based on the Bayesian marginal likelihood.

3 Application to Linear Regression Models

3.1 Criteria

In this section, we derive variable selection criteria for normal linear regression models. First, we consider a collection of candidate models, defined as follows. Let the $n \times p_\omega$ matrix $\boldsymbol{X}_\omega$ consist of all explanatory variables, and assume that $\text{rank}(\boldsymbol{X}_\omega) = p_\omega$. In order to define candidate models using the index set j, suppose that j denotes a subset of $\omega = \{1, \ldots, p_\omega\}$ containing p_j elements (*i.e.*, $p_j = \#(j)$) and that $\boldsymbol{X}_j$ consists of p_j columns of $\boldsymbol{X}_\omega$ indexed by the elements of j. We define the class of candidate models as $\mathcal{J} = \mathcal{P}(\omega)$, namely the power set of ω, where ω denotes the full model. We assume that the true model exists in the class of the candidate models $\mathcal{J}$, which is denoted by j_*. Note that the dimension of the true models is p_{j_*}, which we abbreviate to p_*.

The candidate model j is the linear regression model

$$\boldsymbol{y} = \boldsymbol{X}_j \boldsymbol{\beta}_j + \boldsymbol{\varepsilon}, \tag{7}$$

where $\boldsymbol{y}$ is an $n \times 1$ observation vector of the response variables, $\boldsymbol{X}_j$ is an $n \times p_j$ matrix of the explanatory variables, $\boldsymbol{\beta}_j$ is a $p_j \times 1$ vector of the regression coefficients, and $\boldsymbol{\varepsilon}$ is an $n \times 1$ vector of the random errors. Here, $\boldsymbol{\varepsilon}$ has the distribution $\mathcal{N}_n(\boldsymbol{0}, \sigma^2 \boldsymbol{V})$, where σ^2 is an unknown scalar and $\boldsymbol{V}$ is a known positive definite matrix.

We consider the problem of selecting the explanatory variables, and assume that the true model can be expressed by each candidate model. This is the common assumption used to derive an information criterion. Under this assumption, the true mean of $\boldsymbol{y}$ can be written as

$$E(\boldsymbol{y}) = \boldsymbol{X}_j \boldsymbol{\beta}_j^*,$$

where $\boldsymbol{\beta}_j^*$ is a $p_j \times 1$ vector, the $p_j - p_*$ components of which are exactly 0, and the remaining components are not 0. Hereafter, we omit the model index, j, for notational convenience. Furthermore, we abbreviate $\boldsymbol{\beta}_j^*$ as $\boldsymbol{\beta}$.

Now, we construct the variable selection criteria for the regression model (7), which has the form (5). We consider the following two situations.

[i] A normal prior for $\boldsymbol{\beta}$. We first assume a normal prior distribution for $\boldsymbol{\beta}$,

$$\pi(\boldsymbol{\beta}|\sigma^2) \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{W}),$$

where $\boldsymbol{W}$ is a $p \times p$ matrix, suitably chosen with full rank. Examples of $\boldsymbol{W}$ are $\boldsymbol{W} = (\lambda \boldsymbol{X}^t \boldsymbol{X})^{-1}$ for $\lambda > 0$, when $\boldsymbol{V}$ is the identity matrix, as introduced by Zellner (1986), or more simply, $\boldsymbol{W} = \lambda^{-1} \boldsymbol{I}_p$. For the moment, we assume that λ is known. We discuss how to determine it in Section 3.2. Because the likelihood is $f(\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2) \sim \mathcal{N}(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{V})$, the marginal likelihood function is

$$
\begin{aligned}
f_\pi(\boldsymbol{y}|\sigma^2) &= \int f(\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2) \pi(\boldsymbol{\beta}|\sigma^2) \mathrm{d}\boldsymbol{\beta} \\
&= (2\pi\sigma^2)^{-n/2} \cdot |\boldsymbol{V}|^{-1/2} \cdot |\boldsymbol{W}\boldsymbol{X}^t \boldsymbol{V}^{-1}\boldsymbol{X} + \boldsymbol{I}_p|^{-1/2} \cdot \exp\left\{-\boldsymbol{y}^t \boldsymbol{A}\boldsymbol{y}/(2\sigma^2)\right\},
\end{aligned}
$$

where $\boldsymbol{A} = \boldsymbol{V}^{-1} - \boldsymbol{V}^{-1}\boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{V}^{-1}\boldsymbol{X} + \boldsymbol{W}^{-1})^{-1}\boldsymbol{X}^t\boldsymbol{V}^{-1}$. Note that $\boldsymbol{A} = (\boldsymbol{V} + \boldsymbol{B})^{-1}$ for $\boldsymbol{B} = \boldsymbol{X}\boldsymbol{W}\boldsymbol{X}^t$; that is $f_\pi(\boldsymbol{y}|\sigma^2) \sim \mathcal{N}(\boldsymbol{0}, \sigma^2(\boldsymbol{V} + \boldsymbol{B}))$. Then, we take the predictive density as $\hat{f}(\widetilde{\boldsymbol{y}}; \boldsymbol{y}) = f_\pi(\widetilde{\boldsymbol{y}}|\hat{\sigma}^2)$, and the information (4) can be written as

$$I_{\pi,1}(\boldsymbol{\omega}) = E_{\boldsymbol{\omega}}\left[n\log(2\pi\hat{\sigma}^2) + \log|\boldsymbol{V}| + \log|\boldsymbol{W}\boldsymbol{X}^t\boldsymbol{V}^{-1}\boldsymbol{X} + \boldsymbol{I}_p| + \widetilde{\boldsymbol{y}}^t\boldsymbol{A}\widetilde{\boldsymbol{y}}/\hat{\sigma}^2\right], \qquad (8)$$

where $\hat{\sigma}^2 = \boldsymbol{y}^t\boldsymbol{P}\boldsymbol{y}/n$, $\boldsymbol{P} = \boldsymbol{V}^{-1} - \boldsymbol{V}^{-1}\boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{V}^{-1}$, and $E_{\boldsymbol{\omega}}$ denotes the expectation with respect to the distribution of $f(\widetilde{\boldsymbol{y}}, \boldsymbol{y}|\boldsymbol{\beta}, \sigma^2) = f(\widetilde{\boldsymbol{y}}|\boldsymbol{\beta}, \sigma^2)f(\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2)$ for $\boldsymbol{\omega} = (\boldsymbol{\beta}^t, \sigma^2)^t$. Note that $\boldsymbol{\beta}$ is the parameter of interest, and σ^2 is the nuisance parameter corresponding to $\boldsymbol{\theta}$ in the previous section. Then, we propose the following information criterion.

Proposition 1 *The information $I_{\pi,1}(\boldsymbol{\omega})$ in (8) is unbiasedly estimated by the information criterion*

$$\mathrm{IC}_{\pi,1} = -2\log\{f_\pi(\boldsymbol{y}|\hat{\sigma}^2)\} + \frac{2n}{n - p - 2}, \qquad (9)$$

where

$$-2\log\{f_\pi(\boldsymbol{y}|\hat{\sigma}^2)\} = n\log(2\pi\hat{\sigma}^2) + \log|\boldsymbol{V}| + \log|\boldsymbol{W}\boldsymbol{X}^t\boldsymbol{V}^{-1}\boldsymbol{X} + \boldsymbol{I}_p| + \boldsymbol{y}^t\boldsymbol{A}\boldsymbol{y}/\hat{\sigma}^2;$$

that is, $E_{\boldsymbol{\omega}}(\mathrm{IC}_{\pi,1}) = I_{\pi,1}(\boldsymbol{\omega})$.

If $n^{-1}\boldsymbol{W}^{1/2}\boldsymbol{X}^t\boldsymbol{V}^{-1}\boldsymbol{X}\boldsymbol{W}^{1/2}$ converges to a $p \times p$ positive definite matrix as $n \to \infty$, $\log|\boldsymbol{W}\boldsymbol{X}^t\boldsymbol{V}^{-1}\boldsymbol{X} + \boldsymbol{I}_p|$ can be approximated as $p\log n$, which is the penalty term of the BIC. In that case, $\mathrm{IC}_{\pi,1}$ is approximately expressed as

$$\mathrm{IC}_{\pi,1}^* = n\log(2\pi\hat{\sigma}^2) + \log|\boldsymbol{V}| + p\log n + 2 + \boldsymbol{y}^t\boldsymbol{A}\boldsymbol{y}/\hat{\sigma}^2$$

when n is large.

Note that only the first term of $\mathrm{IC}_{\pi,1}$ can work as a variable selection criterion because $f_\pi(\boldsymbol{y}|\sigma^2)$ is the Bayesian marginal likelihood. The difference between them is

$$\frac{2n}{n - p - 2} = 2 + \frac{2(p + 2)}{n} + O(n^{-2}).$$

In other words, $\mathrm{IC}_{\pi,1}$ has a slight additional penalty, of order n^{-1}. We compare the performance of the criteria using simulations in Section 4.

Alternatively, the KL risk $r(\boldsymbol{\psi}; \hat{f})$ in (6) can be used to evaluate the risk of the predictive density $f_\pi(\widetilde{\boldsymbol{y}}|\hat{\sigma}^2)$ because the prior distribution is proper. Then, the resulting criterion is

$$\mathrm{IC}_{\pi,2} = n\log(2\pi\hat{\sigma}^2) + \log|\boldsymbol{V}| + p\log n + p, \qquad (10)$$

which is an asymptotically unbiased estimator of $I_{\pi,2}(\sigma^2) = E_\pi[I_{\pi,1}(\boldsymbol{\omega})]$, where E_π denotes the expectation with respect to the prior distribution $\pi(\boldsymbol{\beta}|\sigma^2)$; that is $E_\pi E_{\boldsymbol{\omega}}(\mathrm{IC}_{\pi,2}) \to I_{\pi,2}(\sigma^2)$ as $n \to \infty$. Interestingly, $\mathrm{IC}_{\pi,2}$ is analogous to the criterion proposed by Bozdogan (1987), known as the consistent AIC, who suggested replacing the penalty term $2p$ in the AIC with $p + p\log n$.

[ii] Uniform prior for $\boldsymbol{\beta}$. We next assume a uniform prior for $\boldsymbol{\beta}$, namely $\boldsymbol{\beta} \sim uniform(\mathbb{R}^p)$. Although this is an improper prior distribution, we can obtain the marginal likelihood function formally, as follows:

$$
\begin{aligned}
f_r(\boldsymbol{y}|\sigma^2) &= \int f(\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2) \mathrm{d}\boldsymbol{\beta} \\
&= (2\pi\sigma^2)^{-(n-p)/2} \cdot |\boldsymbol{V}|^{-1/2} \cdot |\boldsymbol{X}^t \boldsymbol{V}^{-1} \boldsymbol{X}|^{-1/2} \cdot \exp\left\{-\boldsymbol{y}^t \boldsymbol{P} \boldsymbol{y}/(2\sigma^2)\right\},
\end{aligned}
$$

which is known as the residual likelihood (Patterson and Thompson, 1971). Then, we take the predictive density as $\hat{f}(\widetilde{\boldsymbol{y}}; \boldsymbol{y}) = f_r(\widetilde{\boldsymbol{y}}|\tilde{\sigma}^2)$, and the information (4) can be written as

$$
I_r(\boldsymbol{\omega}) = E_{\boldsymbol{\omega}}\left[(n-p)\log(2\pi\tilde{\sigma}^2) + \log|\boldsymbol{V}| + \log|\boldsymbol{X}^t \boldsymbol{V}^{-1} \boldsymbol{X}| + \widetilde{\boldsymbol{y}}^t \boldsymbol{P} \widetilde{\boldsymbol{y}}/\tilde{\sigma}^2\right], \tag{11}
$$

where $\tilde{\sigma}^2 = \boldsymbol{y}^t \boldsymbol{P} \boldsymbol{y}/(n-p)$, which is the residual maximum likelihood (REML) estimator of σ^2, based on the residual likelihood $f_r(\boldsymbol{y}|\sigma^2)$. Next, we propose the information criterion.

Proposition 2 *The information $I_r(\boldsymbol{\omega})$ in (11) is unbiasedly estimated by the infomation criterion*

$$
\mathrm{IC}_r = -2\log\{f_r(\boldsymbol{y}|\tilde{\sigma}^2)\} + \frac{2(n-p)}{n-p-2}, \tag{12}
$$

where

$$
-2\log\{f_r(\boldsymbol{y}|\tilde{\sigma}^2)\} = (n-p)\log(2\pi\tilde{\sigma}^2) + \log|\boldsymbol{V}| + \log|\boldsymbol{X}^t \boldsymbol{V}^{-1} \boldsymbol{X}| + \boldsymbol{y}^t \boldsymbol{P} \boldsymbol{y}/\tilde{\sigma}^2;
$$

that is, $E_{\boldsymbol{\omega}}(\mathrm{IC}_r) = I_r(\boldsymbol{\omega})$.

Note that $\boldsymbol{y}^t \boldsymbol{P} \boldsymbol{y}/\tilde{\sigma}^2 = n - p$. If $n^{-1}\boldsymbol{X}^t \boldsymbol{V}^{-1} \boldsymbol{X}$ converges to a $p \times p$ positive definite matrix as $n \to \infty$, $\log|\boldsymbol{X}^t \boldsymbol{V}^{-1} \boldsymbol{X}|$ can be approximated by $p\log n$. Then, we can approximate the criterion as

$$
\mathrm{IC}_r^* = (n-p)\log(2\pi\tilde{\sigma}^2) + \log|\boldsymbol{V}| + p\log n + \frac{(n-p)^2}{n-p-2}, \tag{13}
$$

for large n. Note that IC_r^* is equivalent to the RIC proposed by Shi and Tsai (2002). Noting that $(n-p)^2/(n-p-2) = (n+2) + \{4/(n-p-2) - p\}$, we can see that the difference between IC_r^* and the RIC is $n + 2 - p\log(2\pi\tilde{\sigma}^2)$; that is $\mathrm{IC}_r^* = \mathrm{RIC} + n + 2 - p\log(2\pi\tilde{\sigma}^2)$. Note too that the criterion based on $f_r(\boldsymbol{y}|\sigma^2)$ and $r(\boldsymbol{\psi}; f_r)$ cannot be constructed because its KL risk diverges to infinity.

Remark 1 As discussed in Section 2.2, the marginal likelihood based on an improper prior depends on an arbitrary scalar constant, which can, in general, be problematic when selecting the model. However, our IC_r, or its equivalent RIC, can work as a variable selection criterion. In order to show that, we compare IC_r with the AIC and BIC. When $\boldsymbol{V} = \boldsymbol{I}_n$, the AIC and BIC for the normal linear regression model can be expressed as

$$
\mathrm{IC} = n\log(2\pi) + n\log(n^{-1}\mathrm{RSS}) + n + g(p),
$$

where the first three terms are the likelihood part, and the last term $g(p)$ is the penalty, which depends on p. Here, $g(p) = 2(p + 1)$ for the AIC and $g(p) = p \log(n)$ for the BIC. Then, RSS is the residual sum of squares, defined as $\text{RSS} = \|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2 = n\hat{\sigma}^2$. On the other hand, IC_r in (12) can be rewritten as

$$\text{IC}_r = n \log(2\pi) + n \log(n^{-1}\text{RSS}) + n + h(p),$$

where the first three terms are the same as those of the AIC and BIC, and $h(p)$ is

$$\begin{aligned}
h(p) = {} & p\{\log(n - p) - \log(n^{-1}\text{RSS}) - \log(2\pi) - 1\} + \log|\boldsymbol{X}^t\boldsymbol{X}| - p\log(n) \\
& + n \log\{n/(n - p)\} + 2 + O(n^{-1}).
\end{aligned}$$

Thus, $h(p)$ can represent the penalty for the large model because $\log|\boldsymbol{X}^t\boldsymbol{X}| - p\log(n)$ is asymptotically negligible, and the value in the braces of the first term is positive when n is at least moderately large, noting that $n^{-1}\text{RSS}$ becomes small as p becomes large.

3.2 Typical examples of the linear regression models

In the derivation of the criteria, we assumed that the scaled covariance matrix $\boldsymbol{V}$ of the vector of error terms is known. However, it is often the case that $\boldsymbol{V}$ is unknown, and is some function of the unknown parameter $\boldsymbol{\phi}$, namely $\boldsymbol{V} = \boldsymbol{V}(\boldsymbol{\phi})$. In that case, $\boldsymbol{V}$ in each criterion is replaced with its plug-in estimator $\boldsymbol{V}(\widehat{\boldsymbol{\phi}})$, where $\widehat{\boldsymbol{\phi}}$ is some consistent estimator of $\boldsymbol{\phi}$. This strategy is also used in many other studies, for example in Shi and Tsai (2002), who proposed the RIC. We suggest that the $\boldsymbol{\phi}$ be estimated based on the full model. The scaled covariance matrix $\boldsymbol{W}$ of the prior distribution of $\boldsymbol{\beta}$ is also assumed to be known. In practice, its structure should be specified, and we have to estimate the parameter λ in $\boldsymbol{W}$ from the data. In the same manner as $\boldsymbol{V}$, $\boldsymbol{W}$ in each criterion is replaced with $\boldsymbol{W}(\hat{\lambda})$. Note that λ should be estimated based on each candidate model under consideration, because the structure of $\boldsymbol{W}$ depends on the model. We propose that λ is estimated by maximizing the marginal likelihood $f_\pi(\boldsymbol{y}|\hat{\sigma}^2, \lambda)$, after substituting in the estimate $\hat{\sigma}^2$.

Here, we give three examples for the regression model (7): a regression model with constant variance, a variance components model, and a regression model with ARMA errors. The second and the third models include the unknown parameter in the covariance matrix.

[1] **Regression model with constant variance**. When $\boldsymbol{V} = \boldsymbol{I}_n$, (7) represents a multiple regression model with constant variance. In this model, the scaled covariance matrix $\boldsymbol{V}$ does not contain any unknown parameters.

[2] **Variance components model**. Consider a variance components model (Henderson, 1950), described as follows:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}_2\boldsymbol{v}_2 + \cdots + \boldsymbol{Z}_r\boldsymbol{v}_r + \boldsymbol{\eta}, \tag{14}$$

where $\boldsymbol{Z}_i$ is an $n \times m_i$ matrix with $\boldsymbol{V}_i = \boldsymbol{Z}_i\boldsymbol{Z}_i^t$, $\boldsymbol{v}_i$ is an $m_i \times 1$ random vector with distribution $\mathcal{N}_{m_i}(\boldsymbol{0}, \theta_i\boldsymbol{I}_{m_i})$ for $i \geq 2$, $\boldsymbol{\eta}$ is an $n \times 1$ random vector with $\boldsymbol{\eta} \sim \mathcal{N}_n(\boldsymbol{0}, \boldsymbol{V}_0 + \theta_1\boldsymbol{V}_1)$ for known

$n \times n$ matrices $\boldsymbol{V}_0$ and $\boldsymbol{V}_1$, and $\boldsymbol{\eta}$, $\boldsymbol{v}_2, \ldots, \boldsymbol{v}_r$ are mutually independently distributed. The nested error regression model (NERM) is a special case of a variance components model, given by

$$y_{ik} = \boldsymbol{x}_{ik}^t \boldsymbol{\beta} + v_i + \eta_{ik}, \quad (i = 1, \ldots, m; \ k = 1, \ldots, n_i), \tag{15}$$

where v_i and η_{ik} are mutually independently distributed as $v_i \sim \mathcal{N}(0, \tau^2)$ and $\eta_{ik} \sim \mathcal{N}(0, \sigma^2)$, respectively, and $n = \sum_{i=1}^m n_i$. Note that the NERM in (15) is given by $\theta_1 = \sigma^2$, $\theta_2 = \tau^2$, $\boldsymbol{V}_1 = \boldsymbol{I}_n$ and $\boldsymbol{Z}_2 = \text{diag}\,(\mathbf{1}_{n_1}, \ldots, \mathbf{1}_{n_m})$, where $\mathbf{1}_l$ is the l-dimensional vector of ones, for the variance components model (14). This model is often used for clustered data, where v_i is considered the random effect of the cluster (Battese et al., 1988). For such a model, when we are interested in a specific cluster or in predicting the random effects, an appropriate criterion is the conditional AIC, as proposed by Vaida and Blanchard (2005), which is based on the conditional likelihood given the random effects. However, when we wish to predict the fixed effects, namely $\boldsymbol{x}_{ik}^t \boldsymbol{\beta}$, the NERM can be seen as a linear regression model and the random effects are part of the error term. In other words, we consider $\boldsymbol{\varepsilon} = \boldsymbol{Z}_2 \boldsymbol{v}_2 + \boldsymbol{\eta}$, $\boldsymbol{V} = \boldsymbol{V}(\phi) = \phi \boldsymbol{V}_2 + \boldsymbol{I}_n$ for (7), where $\phi = \tau^2/\sigma^2$ and $\boldsymbol{V}_2 = \boldsymbol{Z}_2 \boldsymbol{Z}_2^t = \text{diag}\,(\boldsymbol{J}_{n_1}, \ldots, \boldsymbol{J}_{n_m})$ for $\boldsymbol{J}_l = \mathbf{1}_l \mathbf{1}_l^t$. In this case, our proposed variable selection procedure is useful.

[3] **Regression model with autoregressive moving average errors**. Consider the regression model (7), assuming the random errors are generated by an $\text{ARMA}(q, r)$ process defined by

$$\varepsilon_i - \phi_1 \varepsilon_{i-1} - \cdots - \phi_q \varepsilon_{i-q} = u_i - \varphi_1 u_{i-1} - \cdots - \varphi_r u_{i-r},$$

where $\{u_i\}$ is a sequence of independent normal random variables, with mean 0 and variance τ^2. A special case of this model is the regression model with AR(1) errors, satisfying $\varepsilon_1 \sim \mathcal{N}(0, \tau^2/(1 - \phi^2))$, $\varepsilon_i = \phi \varepsilon_{i-1} + u_i$, and $u_i \sim \mathcal{N}(0, \tau^2)$ for $i = 2, 3, \ldots, n$. When we define $\sigma^2 = \tau^2/(1 - \phi^2)$, the (i, j)-element of the scaled covariance matrix $\boldsymbol{V}$ in (7) is $\phi^{|i-j|}$.

3.3 Consistency of the criteria

In this subsection, we prove that the proposed criteria exhibit consistency. Our asymptotic framework is that n tends to infinity and the true dimension of the regression coefficients p_* is fixed. Following Shi and Tsai (2002), we first show that the criteria are consistent for the regression model with constant variance and pre-specified $\boldsymbol{W}$. Then, we extend the result to the regression model with a general covariance matrix and the case where $\boldsymbol{W}$ is estimated.

We divide $\mathcal{J}$ into two subsets, $\mathcal{J}_+$ and $\mathcal{J}_-$, where $\mathcal{J}_+ = \{j \in \mathcal{J} : j_* \subseteq j\}$ and $\mathcal{J}_- = \mathcal{J} \backslash \mathcal{J}_+$. Note that the true model j_* is the smallest model in $\mathcal{J}_+$, and that $E(\boldsymbol{y}) = \boldsymbol{X}_{j_*} \boldsymbol{\beta}_{j_*}$, abbreviated to $\boldsymbol{X}_* \boldsymbol{\beta}_*$, where $\boldsymbol{\beta}_*$ is a $p_* \times 1$ vector of the true regression coefficients. Let $\hat{j}$ denote the model selected by some criterion. Following Shi and Tsai (2002), we make the following assumptions:

(A1) $E(\varepsilon_1^4) < \infty$.

(A2) $0 < \liminf_{n \to \infty} \min_{j \in \mathcal{J}} |\boldsymbol{X}_j^t \boldsymbol{X}_j / n|$ and $\limsup_{n \to \infty} \max_{j \in \mathcal{J}} |\boldsymbol{X}_j^t \boldsymbol{X}_j / n| < \infty$.

(A3) $\liminf_{n \to \infty} n^{-1} \inf_{j \in \mathcal{J}_-} \|\boldsymbol{X}_* \boldsymbol{\beta}_* - \boldsymbol{H}_j \boldsymbol{X}_* \boldsymbol{\beta}_*\|^2 > 0$, where $\boldsymbol{H}_j = \boldsymbol{X}_j (\boldsymbol{X}_j^t \boldsymbol{X}_j)^{-1} \boldsymbol{X}_j^t$.

We can now obtain the asymptotic properties of the criteria for the regression model with constant variance.

Theorem 1 *If assumptions* (A1)–(A3) *are satisfied,* $\mathcal{J}_+$ *is not empty, the* ε_i*'s are independent and identically distributed (iid), and* $\boldsymbol{W}_j$ *in the prior distribution of* $\boldsymbol{\beta}_j$ *is pre-specified, then the criteria* $\mathrm{IC}_{\pi,1}$, $\mathrm{IC}_{\pi,1}^*$, $\mathrm{IC}_{\pi,2}$, IC_r, *and* IC_r^* *are consistent; that is* $P(\hat{j} = j_*) \to 1$ *as* $n \to \infty$.

The proof of Theorem 1 is given in Appendix B.

We next consider the regression model with a general covariance structure and the case where $\boldsymbol{W}_j$ is estimated from the data. In this case, $\boldsymbol{V}$ and $\boldsymbol{W}_j$ are replaced with their plug-in estimators $\boldsymbol{V}(\widehat{\boldsymbol{\phi}})$ and $\boldsymbol{W}_j(\hat{\lambda}_j)$, respectively.

Theorem 2 *Assume that* $\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi}_0$ *and* $\hat{\lambda}_j - \lambda_{j,0}$ *tend to 0 in probability as* $n \to \infty$, *for all* $j \in \mathcal{J}$. *In addition, assume that the elements of* $\boldsymbol{V}(\boldsymbol{\phi})$ *and* $\boldsymbol{W}_j(\lambda_j)$ *are continuous functions of* $\boldsymbol{\phi}$ *and* λ_j, *respectively, and that* $\boldsymbol{V}(\boldsymbol{\phi})$ *and* $\boldsymbol{W}_j(\boldsymbol{\lambda}_j)$ *are positive definite in the neighborhood of* $\boldsymbol{\phi}_0$ *and* $\lambda_{j,0}$, *respectively, for all* $j \in \mathcal{J}$. *If assumptions* (A1)–(A3) *are satisfied when* $\boldsymbol{X}_j$ *and* $\boldsymbol{\varepsilon}$ *are replaced with* $\boldsymbol{V}^{-1/2}\boldsymbol{X}_j$ *and* $\boldsymbol{\varepsilon}^* = \boldsymbol{V}^{-1/2}\boldsymbol{\varepsilon}$, *respectively,* $\mathcal{J}_+$ *is not empty and* ε_i^* *are iid. Then, the criteria* $\mathrm{IC}_{\pi,1}$, $\mathrm{IC}_{\pi,1}^*$, $\mathrm{IC}_{\pi,2}$, IC_r, *and* IC_r^* *are consistent.*

For the proof of Theorem 2, we use the same techniques as those used in the proof of Theorem 1.

4 Simulations

In this section, we compare the numerical performance of the proposed criteria, $\mathrm{IC}_{\pi,1}$ and IC_r, with that of conventional criteria, namely the AIC, BIC, DIC, and the marginal likelihood. We consider two regression models: a regression model with constant variance and a regression model with AR(1) errors. These models are taken as examples of the linear model (7) given in Section 3.2. In each simulation, 1000 realizations are generated from (7), with $\boldsymbol{\beta} = (1,1,1,1,0,0,0)^t$ or $\boldsymbol{\beta} = (1,1,0,0,0,0,0,)^t$. That is, the full model is $p_\omega = 7$ and the true model is $p_* = 4$ or $p_* = 2$. All explanatory variables are randomly generated from the standard normal distribution. The signal-to-noise ratio $(\mathrm{SNR} = \{\mathrm{var}(\boldsymbol{x}_i^t\boldsymbol{\beta})/\mathrm{var}(\varepsilon_i)\}^{1/2})$ is controlled at 1, 3, and 5. For the regression model with AR(1) errors, we set the AR parameter as $\phi = 0.5$.

When deriving the criterion $\mathrm{IC}_{\pi,1}$, we set the prior distribution of $\boldsymbol{\beta}$ as $\mathcal{N}_p(\boldsymbol{0}, \sigma^2\lambda^{-1}\boldsymbol{I}_p)$; that is, $\boldsymbol{W} = \lambda^{-1}\boldsymbol{I}_p$. The unknown parameter ϕ in $\boldsymbol{V}$ for the AR(1) model is estimated using the maximum likelihood estimator based on the full model. The hyperparameter λ is estimated by maximizing the marginal likelihood $f_\pi(\boldsymbol{y}|\hat{\sigma}^2, \lambda)$, after substituting in the estimate $\hat{\sigma}^2 = \boldsymbol{y}^t\boldsymbol{P}\boldsymbol{y}/n$ of σ^2. Note that ϕ is estimated based on the full model, while σ^2 and λ are estimated from each candidate model using the plugged-in version of $\boldsymbol{V}(\hat{\phi})$.

As a competitor for $\mathrm{IC}_{\pi,1}$, we consider the criterion that uses $f_\pi(\boldsymbol{y}|\hat{\sigma}^2)$ only, which is the so-called the marginal likelihood commonly used in Bayesian analyses. Another competitor is the DIC, which is also popular in Bayesian analyses. When deriving the DIC, we consider the same prior distribution of $\boldsymbol{\beta}$ as that assumed when deriving $\mathrm{IC}_{\pi,1}$, namely $\boldsymbol{\beta} \sim \mathcal{N}_p(\boldsymbol{0}, \sigma^2\lambda^{-1}\boldsymbol{I}_p)$. In fairness to the other criteria, we take σ^2 as an unknown parameter and do not assume a prior distribution for the derivation of the DIC. Let $D(\boldsymbol{\beta}) = -2\log\{f(\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2)\}$. When σ^2 is known, the DIC is

$$\mathrm{DIC}(\sigma^2) = 2E_{\boldsymbol{\beta}|\boldsymbol{y}}[D(\boldsymbol{\beta})] - D(\widetilde{\boldsymbol{\beta}}),$$

where $E_{\boldsymbol{\beta}|\boldsymbol{y}}$ denotes the expectation with respect to the conditional distribution of $\boldsymbol{\beta}$, given $\boldsymbol{y}$ and $\widetilde{\boldsymbol{\beta}} = E_{\boldsymbol{\beta}|\boldsymbol{y}}(\boldsymbol{\beta})$. Because $\boldsymbol{\beta}|\boldsymbol{y} \sim \mathcal{N}_p(\widetilde{\boldsymbol{\beta}}, \sigma^2(\boldsymbol{X}^t\boldsymbol{V}^{-1}\boldsymbol{X} + \boldsymbol{W}^{-1})^{-1})$ for $\widetilde{\boldsymbol{\beta}} = (\boldsymbol{X}^t\boldsymbol{V}^{-1}\boldsymbol{X} + \boldsymbol{W}^{-1})^{-1}\boldsymbol{X}^t\boldsymbol{V}^{-1}\boldsymbol{y}$, the first term of the DIC is

$$2E_{\boldsymbol{\beta}|\boldsymbol{y}}E[D(\boldsymbol{\beta})] = 2\mathrm{tr}\left[\boldsymbol{X}^t\boldsymbol{V}^{-1}\boldsymbol{X}\{\sigma^2(\boldsymbol{X}^t\boldsymbol{V}^{-1}\boldsymbol{X} + \boldsymbol{W}^{-1})^{-1} + \widetilde{\boldsymbol{\beta}}\widetilde{\boldsymbol{\beta}}^t\}\right]/\sigma^2 - 4\boldsymbol{y}^t\boldsymbol{V}^{-1}\boldsymbol{X}\widetilde{\boldsymbol{\beta}}/\sigma^2$$
$$+ \text{(the term which is irrelevant to the model)},$$

and the second term is $D(\widetilde{\boldsymbol{\beta}}) = (\boldsymbol{y} - \boldsymbol{X}\widetilde{\boldsymbol{\beta}})^t\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X}\widetilde{\boldsymbol{\beta}})/\sigma^2 +$ (the term that is irrelevant to the model). Then, we use $\mathrm{DIC}(\hat{\sigma}^2)$, where $\hat{\sigma}^2$ is substituted into $\mathrm{DIC}(\sigma^2)$.

We consider all subsets of the full model as the class of the candidate models. The performance of the criteria is measured as the number of simulations that select the true model and the prediction error of the selected model, based on the quadratic loss; that is $\|\boldsymbol{X}_{\hat{j}}\widehat{\boldsymbol{\beta}}_{\hat{j}} - \boldsymbol{X}_*\boldsymbol{\beta}_*\|^2/n$, where $\widehat{\boldsymbol{\beta}}_{\hat{j}} = (\boldsymbol{X}_{\hat{j}}^t\boldsymbol{V}(\hat{\phi})^{-1}\boldsymbol{X}_{\hat{j}})^{-1}\boldsymbol{X}_{\hat{j}}^t\boldsymbol{V}(\hat{\phi})^{-1}\boldsymbol{y}$ (i.e., the GLS estimator).

First, we confirm that $\mathrm{IC}_{\pi,1}$, IC_r, and the BIC are consistent. Figure 1 shows the number of simulations that select the true model among 1000 realizations of the regression model with constant variance. The resuls show that each criterion is consistent. When the data are noisy (i.e., the SNR is weak), IC_r does not perform as well as $\mathrm{IC}_{\pi,1}$ and the BIC in terms of selecting the true model. Although we omit the detail, the results for the AR(1) model are similar to those of the model with constant variance.

Next, we investigate the performance of the criteria in terms of the prediction error. Tables 1–4 show the mean value of the prediction error among 1000 simulations. In each case, we put an asterisk at the minimum value of the prediction error. In many cases, $\mathrm{IC}_{\pi,1}$ performs best, except when the sample size is small and the SNR is weak. Although the performance of the marginal likelihood (ML) is similar to $\mathrm{IC}_{\pi,1}$, that of $\mathrm{IC}_{\pi,1}$ is slightly better. For the noisy and small sample size data, which is the exception to the good performance of $\mathrm{IC}_{\pi,1}$, IC_r performs very well, except for the regression with constant variance with $p_* = 2$, shown in Table 2. In this case, the result in Figure 1 show that IC_r cannot select the true model. In this case, IC_r tends to select larger models. Shi and Tsai (2002) also pointed out that the RIC, which is equivalent to IC_r, does not perform well when the sample size is small and the SNR is weak. On the other hand, IC_r performs best in terms of the prediction.

Figure 1: The number of simulations that select the true model by the criteria in 1000 realizations of the regression model with constant variance. The left three figures are the result for $p_* = 2$, and the right three figures are for $p_* = 4$.
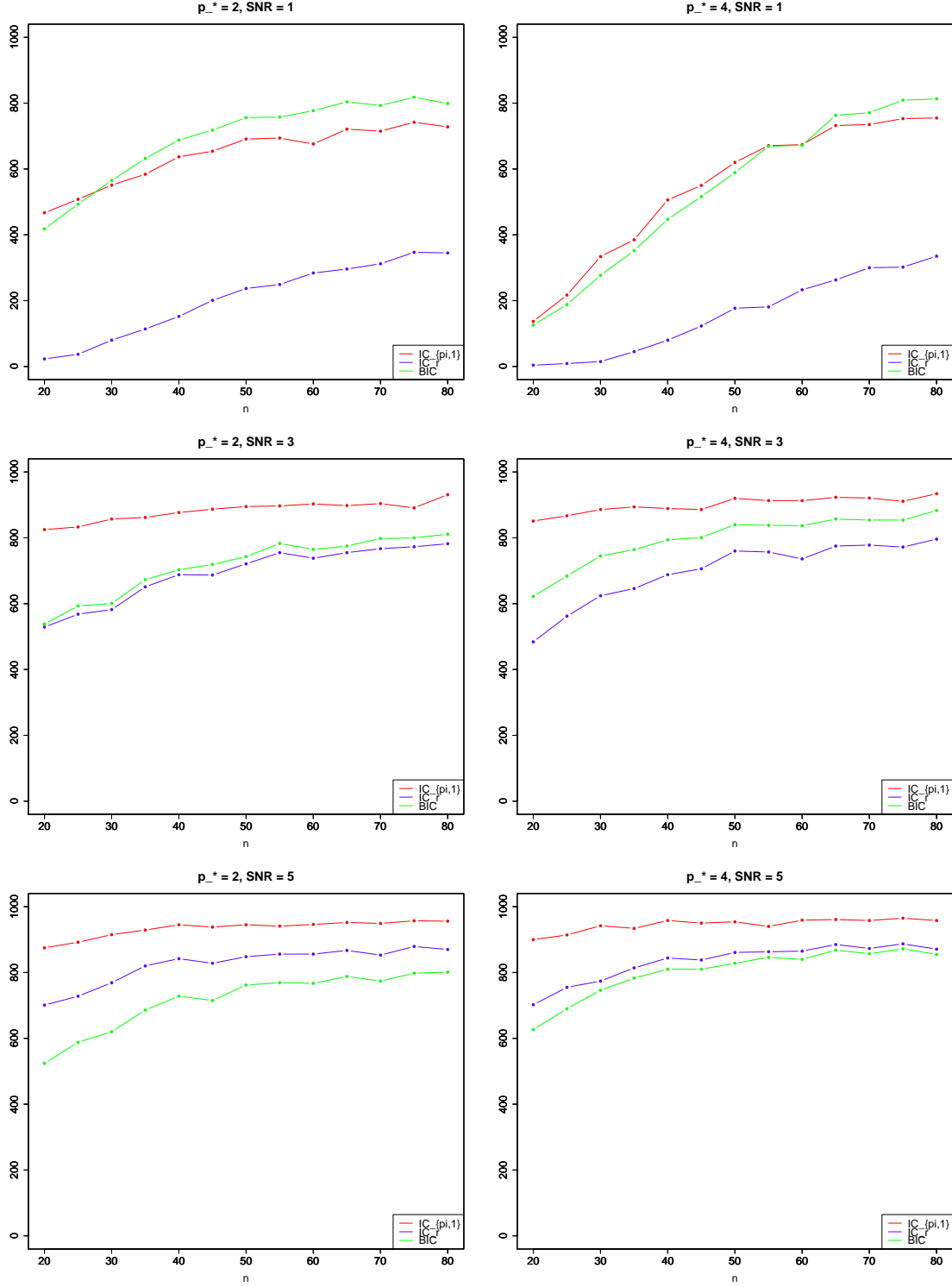
Table 1: The mean of the prediction error among 1000 simulations for the regression model with constant variance. The true model is $p_* = 4$.

	SNR	$IC_{\pi,1}$	IC_r	AIC	BIC	DIC	ML
	1	1.67	*1.41	1.56	1.71	1.42	1.63
$n = 20$	3	*0.105	0.130	0.132	0.122	0.152	0.106
	5	*0.0373	0.0436	0.0492	0.0456	0.0562	0.0377
	1	0.721	*0.698	0.699	0.850	0.702	0.713
$n = 40$	3	*0.0516	0.0598	0.0643	0.0560	0.0764	0.0520
	5	*0.0167	0.0189	0.0224	0.0194	0.0268	0.0168
	1	*0.274	0.326	0.299	0.275	0.348	0.275
$n = 80$	3	*0.0241	0.0274	0.0314	0.0254	0.0380	0.0241
	5	*0.00877	0.00972	0.0120	0.00984	0.0144	0.00878

Table 2: The mean of the prediction error among 1000 simulations for the regression model with constant variance. The true model is $p_* = 2$.

	SNR	$IC_{\pi,1}$	IC_r	AIC	BIC	DIC	ML
	1	0.524	0.687	0.604	0.561	0.696	*0.522
$n = 20$	3	*0.0337	0.0486	0.0577	0.0484	0.0757	0.0347
	5	*0.0116	0.0151	0.0219	0.0185	0.0281	0.0117
	1	*0.207	0.304	0.261	0.201	0.342	0.211
$n = 40$	3	*0.0155	0.0206	0.0279	0.0203	0.0382	0.0157
	5	*0.00476	0.00590	0.00969	0.00699	0.0134	0.00477
	1	*0.0920	0.131	0.1268	0.0831	0.170	0.0926
$n = 80$	3	*0.00683	0.00896	0.01348	0.00853	0.01899	*0.00683
	5	*0.00242	0.00296	0.00516	0.00333	0.00720	0.00242

Table 3: The mean of the prediction error among 1000 simulations for AR(1) with $\phi = 0.5$. The true model is $p_* = 4$.

	SNR	$IC_{\pi,1}$	IC_r	AIC	BIC	DIC	ML
	1	0.341	*0.296	0.374	0.449	0.360	0.334
$n = 40$	3	*0.0223	0.0251	0.0312	0.0268	0.0375	0.0224
	5	*0.00729	0.00805	0.0108	0.00895	0.0133	0.00730
	1	*0.179	0.184	0.206	0.252	0.231	0.179
$n = 60$	3	*0.0146	0.0160	0.0199	0.0163	0.0243	*0.0146
	5	*0.00494	0.00527	0.00722	0.00583	0.00885	0.00496
	1	*0.119	0.131	0.143	0.154	0.174	0.119
$n = 80$	3	*0.0109	0.0118	0.0151	0.0121	0.0186	*0.0109
	5	*0.00390	0.00412	0.00551	0.00444	0.00673	*0.00390

Table 4: The mean of the prediction error among 1000 simulations for AR(1) with $\phi = 0.5$. The true model is $p_* = 2$.

	SNR	$IC_{\pi,1}$	IC_r	AIC	BIC	DIC	ML
	1	0.341	*0.296	0.374	0.449	0.360	0.334
$n = 40$	3	*0.0223	0.0251	0.0312	0.0268	0.0375	0.0224
	5	*0.00729	0.00805	0.01079	0.00895	0.0133	0.00730
	1	*0.179	0.184	0.206	0.252	0.231	0.179
$n = 60$	3	*0.0146	0.0160	0.0199	0.0163	0.0243	*0.0146
	5	*0.00494	0.00527	0.00722	0.00583	0.00885	0.00496
	1	*0.119	0.131	0.143	0.154	0.174	0.119
$n = 80$	3	*0.0109	0.0118	0.0151	0.0121	0.0186	*0.0109
	5	*0.00390	0.00412	0.00551	0.00444	0.00673	*0.00390

5 Concluding Remarks

We have derived variable selection criteria for normal linear regression models relative to the frequentist KL risk of the predictive density, based on the Bayesian marginal likelihood. We have proved the consistency of the criteria and, using simulations, have shown that they perform well in terms of the prediction.

Although our theoretical approach is general, the derivation of the criterion depends on the normal distribution. If we assume a conjugate prior distribution for the parameter of interest when deriving the criterion, it is easy to extend our approach to other models. However, for the class of generalized linear models, which includes the Poisson and the logistic regression models, it is difficult to consider a prior distribution where the marginal likelihood can be evaluated analytically. In such models, we have to rely on some computational method, which we leave for future research.

Variable selection for the mixed effects models, such as the variance components model (14) in Section 3.2, is another important problem. As discussed, it is appropriate to consider the KL divergence based on the conditional density, given the random effects, when the objective is to predict the random effects, as in the conditional AIC (cAIC). An extension of our approach to the cAIC-type criterion is also left to future research.

A   Derivations of the Criteria

In this section, we show the derivations of the criteria. To this end, we first obtain the following lemma, which was shown in Section A.2 of Srivastava and Kubokawa (2010).

Lemma 1 *Assume that $\boldsymbol{C}$ is an $n \times n$ symmetric matrix, $\boldsymbol{M}$ is an idempotent matrix of rank p, and that $\boldsymbol{u} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$. Then,*

$$E\left[\frac{\boldsymbol{u}^t \boldsymbol{C} \boldsymbol{u}}{\boldsymbol{u}^t(\boldsymbol{I}_n - \boldsymbol{M})\boldsymbol{u}}\right] = \frac{\operatorname{tr}(\boldsymbol{C})}{n - p - 2} - \frac{2\operatorname{tr}[\boldsymbol{C}(\boldsymbol{I}_n - \boldsymbol{M})]}{(n - p)(n - p - 2)}.$$

A.1   Derivation of $\mathrm{IC}_{\pi,1}$ in (9)

It is sufficient to show that the bias correction $\Delta_{\pi,1} = I_{\pi,1}(\boldsymbol{\omega}) - E_{\boldsymbol{\omega}}[-2\log\{f_\pi(\boldsymbol{y}|\hat{\sigma}^2)\}]$ is $2n/(n - p - 2)$, where $I_{\pi,1}(\boldsymbol{\omega})$ is given by (8). It follows that

$$\begin{aligned}
\Delta_{\pi,1} &= E_{\boldsymbol{\omega}}(\widetilde{\boldsymbol{y}}^t \boldsymbol{A} \widetilde{\boldsymbol{y}}/\hat{\sigma}^2) - E_{\boldsymbol{\omega}}(\boldsymbol{y}^t \boldsymbol{A} \boldsymbol{y}/\hat{\sigma}^2) \\
&= E_{\boldsymbol{\omega}}(\widetilde{\boldsymbol{y}}^t \boldsymbol{A} \widetilde{\boldsymbol{y}}) \cdot E_{\boldsymbol{\omega}}(1/\hat{\sigma}^2) - E_{\boldsymbol{\omega}}(\boldsymbol{y}^t \boldsymbol{A} \boldsymbol{y}/\hat{\sigma}^2).
\end{aligned}$$

First,

$$\begin{aligned}
E_{\boldsymbol{\omega}}(\widetilde{\boldsymbol{y}}^t \boldsymbol{A} \widetilde{\boldsymbol{y}}) &= E_{\boldsymbol{\omega}}[(\widetilde{\boldsymbol{y}} - \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{X}\boldsymbol{\beta})^t \boldsymbol{A} (\widetilde{\boldsymbol{y}} - \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{X}\boldsymbol{\beta})] \\
&= \sigma^2 \operatorname{tr}(\boldsymbol{A}\boldsymbol{V}) + \boldsymbol{\beta}^t \boldsymbol{X}^t \boldsymbol{A} \boldsymbol{X} \boldsymbol{\beta}.
\end{aligned} \tag{16}$$

Second, noting that $n\hat{\sigma}^2 = \boldsymbol{y}^t \boldsymbol{P} \boldsymbol{y} = \sigma^2 \boldsymbol{u}^t(\boldsymbol{I}_n - \boldsymbol{M})\boldsymbol{u}$ for

$$\begin{aligned}
\boldsymbol{u} &= \boldsymbol{V}^{-1/2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})/\sigma, \\
\boldsymbol{M} &= \boldsymbol{I}_n - \boldsymbol{V}^{-1/2}\boldsymbol{X}(\boldsymbol{X}^t \boldsymbol{V}^{-1} \boldsymbol{X})^{-1}\boldsymbol{X}^t \boldsymbol{V}^{-1/2},
\end{aligned} \tag{17}$$

and that $\boldsymbol{P}\boldsymbol{X} = \boldsymbol{0}$, we obtain

$$\begin{aligned}
E_{\boldsymbol{\omega}}(1/\hat{\sigma}^2) &= nE_{\boldsymbol{\omega}}\left(\frac{1}{\boldsymbol{y}^t \boldsymbol{P} \boldsymbol{y}}\right) = nE_{\boldsymbol{\omega}}\left[\frac{1}{\sigma^2 \boldsymbol{u}^t(\boldsymbol{I}_n - \boldsymbol{M})\boldsymbol{u}}\right] \\
&= \frac{n}{\sigma^2(n - p - 2)}.
\end{aligned} \tag{18}$$

Finally,

$$\begin{aligned}
E_{\boldsymbol{\omega}}(\boldsymbol{y}^t \boldsymbol{A} \boldsymbol{y}/\hat{\sigma}^2) &= nE_{\boldsymbol{\omega}}\left(\frac{\boldsymbol{y}^t \boldsymbol{A} \boldsymbol{y}}{\boldsymbol{y}^t \boldsymbol{P} \boldsymbol{y}}\right) = nE_{\boldsymbol{\omega}}\left[\frac{\sigma^2 \boldsymbol{u}^t \boldsymbol{V}^{1/2} \boldsymbol{A} \boldsymbol{V}^{1/2} \boldsymbol{u} + \boldsymbol{\beta}^t \boldsymbol{X}^t \boldsymbol{A} \boldsymbol{X} \boldsymbol{\beta}}{\sigma^2 \boldsymbol{u}^t(\boldsymbol{I}_n - \boldsymbol{M})\boldsymbol{u}}\right] \\
&= n \times \left\{\frac{\operatorname{tr}(\boldsymbol{A}\boldsymbol{V})}{n - p - 2} - \frac{2\operatorname{tr}(\boldsymbol{A}\boldsymbol{V}\boldsymbol{P}\boldsymbol{V})}{(n - p)(n - p - 2)} + \frac{\boldsymbol{\beta}^t \boldsymbol{X}^t \boldsymbol{A} \boldsymbol{X} \boldsymbol{\beta}}{\sigma^2(n - p - 2)}\right\}.
\end{aligned} \tag{19}$$

The latter equation derives from Lemma 1. Combining (16), (18), and (19), we have

$$\Delta_{\pi,1} = \frac{2n \cdot \operatorname{tr}(\boldsymbol{A}\boldsymbol{V}\boldsymbol{P}\boldsymbol{V})}{(n - p)(n - p - 2)}.$$

Therefore,

$$
\begin{aligned}
\mathrm{tr}\,(\boldsymbol{AVPV}) &= \mathrm{tr}\,\{(\boldsymbol{V}+\boldsymbol{B})^{-1}(\boldsymbol{V}+\boldsymbol{B}-\boldsymbol{B})\boldsymbol{PV}\} \\
&= \mathrm{tr}\,(\boldsymbol{PV}) - \mathrm{tr}\,\{(\boldsymbol{V}+\boldsymbol{B})^{-1}\boldsymbol{BPV}\} \\
&= \mathrm{tr}\,(\boldsymbol{I}_n - \boldsymbol{M}) = n - p,
\end{aligned}
\tag{20}
$$

because $\boldsymbol{BP} = \boldsymbol{XWX}^t\boldsymbol{P} = \boldsymbol{0}$. Then, we obtain $\Delta_{\pi,1} = 2n/(n-p-2)$. $\square$

A.2   Derivation of $\mathrm{IC}_{\pi,2}$ in (10)

From the fact that $E_{\boldsymbol{\omega}}(\mathrm{IC}_{\pi,1}) = I_{\pi,1}(\boldsymbol{\omega})$ and that $E_\pi E_{\boldsymbol{\omega}}(\mathrm{IC}_{\pi,1}) = E_\pi[I_{\pi,1}(\boldsymbol{\omega})] = I_{\pi,2}(\sigma^2)$, it suffices to show that $E_\pi E_{\boldsymbol{\omega}}(\mathrm{IC}_{\pi,1})$ is approximated by

$$
\begin{aligned}
E_\pi E_{\boldsymbol{\omega}}(\mathrm{IC}_{\pi,1}) &\approx E_\pi E_{\boldsymbol{\omega}}[n\log(2\pi\hat{\sigma}^2) + \log|\boldsymbol{V}| + p\log n + 2 + \boldsymbol{y}^t\boldsymbol{Ay}/\hat{\sigma}^2] \\
&\approx E_\pi E_{\boldsymbol{\omega}}[n\log(2\pi\hat{\sigma}^2) + \log|\boldsymbol{V}| + p\log n + p] + (n+2) = E_\pi E_{\boldsymbol{\omega}}(\mathrm{IC}_{\pi,2}) + (n+2),
\end{aligned}
$$

when n is large. Note that $n+2$ is irrelevant to the model. It follows that

$$
\begin{aligned}
&E_{\boldsymbol{\omega}}\left(\frac{\boldsymbol{y}^t\boldsymbol{Ay}}{\hat{\sigma}^2}\right) \\
&= n \times E_{\boldsymbol{\omega}}\left[\frac{\boldsymbol{y}^t\{\boldsymbol{V}^{-1} - \boldsymbol{V}^{-1}\boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{V}^{-1}\boldsymbol{X} + \boldsymbol{W}^{-1})^{-1}\boldsymbol{X}^t\boldsymbol{V}^{-1}\}\boldsymbol{y}}{\boldsymbol{y}^t\{\boldsymbol{V}^{-1} - \boldsymbol{V}^{-1}\boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{V}^{-1}\}\boldsymbol{y}}\right] \\
&= n + n \times E_{\boldsymbol{\omega}}\left[\frac{\boldsymbol{y}^t\boldsymbol{V}^{-1}\boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{V}^{-1}\boldsymbol{X} + \boldsymbol{W}^{-1})^{-1}\boldsymbol{W}^{-1}(\boldsymbol{X}^t\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{V}^{-1}\boldsymbol{y}}{\boldsymbol{y}^t\{\boldsymbol{V}^{-1} - \boldsymbol{V}^{-1}\boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{V}^{-1}\}\boldsymbol{y}}\right] \\
&= n + \frac{n}{\sigma^2(n-p-2)} \times E_{\boldsymbol{\omega}}\left[\boldsymbol{y}^t\boldsymbol{V}^{-1}\boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{V}^{-1}\boldsymbol{X} + \boldsymbol{W}^{-1})^{-1}\boldsymbol{W}^{-1}(\boldsymbol{X}^t\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{V}^{-1}\boldsymbol{y}\right] \\
&= n + \frac{n}{\sigma^2(n-p-2)} \times \Big[\sigma^2 \cdot \mathrm{tr}\,\{(\boldsymbol{X}^t\boldsymbol{V}^{-1}\boldsymbol{X} + \boldsymbol{W}^{-1})^{-1}\boldsymbol{W}^{-1}\} \\
&\quad + \boldsymbol{\beta}^t\boldsymbol{X}^t\boldsymbol{V}^{-1}\boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{V}^{-1}\boldsymbol{X} + \boldsymbol{W}^{-1})^{-1}\boldsymbol{W}^{-1}\boldsymbol{\beta}\Big],
\end{aligned}
$$

and that

$$
E_\pi[\boldsymbol{\beta}^t\boldsymbol{X}^t\boldsymbol{V}^{-1}\boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{V}^{-1}\boldsymbol{X} + \boldsymbol{W}^{-1})^{-1}\boldsymbol{W}^{-1}\boldsymbol{\beta}] = \sigma^2 \cdot \mathrm{tr}\,[\boldsymbol{X}^t\boldsymbol{V}^{-1}\boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{V}^{-1}\boldsymbol{X} + \boldsymbol{W}^{-1})^{-1}].
$$

If $n^{-1}\boldsymbol{X}^t\boldsymbol{V}^{-1}\boldsymbol{X}$ converges to a $p \times p$ positive definite matrix as $n \to \infty$, $\mathrm{tr}\,[(\boldsymbol{X}^t\boldsymbol{V}^{-1}\boldsymbol{X} + \boldsymbol{W}^{-1})^{-1}\boldsymbol{W}^{-1}] \to 0$ and $\mathrm{tr}\,[\boldsymbol{X}^t\boldsymbol{V}^{-1}\boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{V}^{-1}\boldsymbol{X} + \boldsymbol{W}^{-1})^{-1}] \to p$. Then, we have $E_\pi E_{\boldsymbol{\omega}}(\boldsymbol{y}^t\boldsymbol{Ay}/\hat{\sigma}^2 - n) \to p$, which we want to show. $\square$

A.3   Derivation of IC_r in (12)

We show that the bias correction $\Delta_r = I_r(\boldsymbol{\omega}) - E_{\boldsymbol{\omega}}[-2\log\{f_r(\boldsymbol{y}|\tilde{\sigma}^2)\}]$ is $2(n-p)/(n-p-2)$, where $I_r(\boldsymbol{\omega})$ is given by (11). Then,

$$
\begin{aligned}
\Delta_r &= E_{\boldsymbol{\omega}}(\widetilde{\boldsymbol{y}}^t\boldsymbol{P}\widetilde{\boldsymbol{y}}/\tilde{\sigma}^2) - E_{\boldsymbol{\omega}}(\boldsymbol{y}^t\boldsymbol{Py}/\tilde{\sigma}^2) \\
&= E_{\boldsymbol{\omega}}(\widetilde{\boldsymbol{y}}^t\boldsymbol{P}\widetilde{\boldsymbol{y}}) \cdot E_{\boldsymbol{\omega}}(1/\tilde{\sigma}^2) - (n-p).
\end{aligned}
$$

Since $E_{\boldsymbol{\omega}}(\widetilde{\boldsymbol{y}}^t \boldsymbol{P} \widetilde{\boldsymbol{y}}) = (n-p)\sigma^2$ and $E_{\boldsymbol{\omega}}(1/\widetilde{\sigma}^2) = (n-p)/\{\sigma^2(n-p-2)\}$, we have $\Delta_r = 2(n-p)/(n-p-2)$. $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\quad\square$

B Proof of Theorem 1

We only prove the consistency of $\text{IC}_{\pi,1}$. The proof of the consistency of the other criteria can be shown in the same manner. Because we have

$$P(\hat{\jmath} = j) \leq P\{\text{IC}_{\pi,1}(j) < \text{IC}_{\pi,1}(j_*)\}$$

for any $j \in \mathcal{J} \setminus \{j_*\}$, it suffices to show that $P\{\text{IC}_{\pi,1}(j) < \text{IC}_{\pi,1}(j_*)\} \to 0$, or equivalently, that $P\{\text{IC}_{\pi,1}(j) - \text{IC}_{\pi,1}(j_*) > 0\} \to 1$ as $n \to \infty$ for $j \in \mathcal{J} \setminus \{j_*\}$. When $\boldsymbol{V} = \boldsymbol{I}_n$, we obtain

$$\text{IC}_{\pi,1}(j) - \text{IC}_{\pi,1}(j_*) = I_1 + I_2 + I_3,$$

where

$$
\begin{aligned}
I_1 &= n\log(\hat{\sigma}_j^2/\hat{\sigma}_*^2) + \boldsymbol{y}^t \boldsymbol{A}_j \boldsymbol{y}/\hat{\sigma}_j^2 - \boldsymbol{y}^t \boldsymbol{A}_* \boldsymbol{y}/\hat{\sigma}_*^2, \\
I_2 &= \log|\boldsymbol{X}_j^t \boldsymbol{X}_j + \boldsymbol{W}_j^{-1}| - \log|\boldsymbol{X}_*^t \boldsymbol{X}_* + \boldsymbol{W}_*^{-1}|, \\
I_3 &= \log\{|\boldsymbol{W}_j|/|\boldsymbol{W}_*|\} + \frac{2n}{n-p_j-2} - \frac{2n}{n-p_*-2},
\end{aligned}
$$

for $\hat{\sigma}_j^2 = \boldsymbol{y}^t(\boldsymbol{I}_n - \boldsymbol{H}_j)\boldsymbol{y}/n$, $\hat{\sigma}_*^2 = \hat{\sigma}_{j_*}^2$, $\boldsymbol{A}_j = \boldsymbol{I}_n - \boldsymbol{X}_j(\boldsymbol{X}_j^t \boldsymbol{X}_j + \boldsymbol{W}_j^{-1})^{-1}\boldsymbol{X}_j^t$, $\boldsymbol{A}_{j_*} = \boldsymbol{A}_*$, and $\boldsymbol{W}_* = \boldsymbol{W}_{j_*}$. We evaluate the asymptotic behaviors of I_1, I_2, and I_3 for $j \in \mathcal{J}_-$, and $j \in \mathcal{J}_+ \setminus \{j_0\}$, separately.

[Case of $j \in \mathcal{J}_-$]. First, we evaluate I_1. We decompose $I_1 = I_{11} + I_{12}$, where $I_{11} = n\log(\hat{\sigma}_j^2/\hat{\sigma}_*^2)$ and $I_{12} = \boldsymbol{y}^t \boldsymbol{A}_j \boldsymbol{y}/\hat{\sigma}_j^2 - \boldsymbol{y}^t \boldsymbol{A}_* \boldsymbol{y}/\hat{\sigma}_*^2$. It follows that

$$
\begin{aligned}
\hat{\sigma}_j^2 - \hat{\sigma}_*^2 &= (\boldsymbol{X}_* \boldsymbol{\beta}_* + \boldsymbol{\varepsilon})^t(\boldsymbol{I}_n - \boldsymbol{H}_j)(\boldsymbol{X}_* \boldsymbol{\beta}_* + \boldsymbol{\varepsilon})/n - \boldsymbol{\varepsilon}^t(\boldsymbol{I}_n - \boldsymbol{H}_*)\boldsymbol{\varepsilon}/n \\
&= \|\boldsymbol{X}_* \boldsymbol{\beta}_* - \boldsymbol{H}_j \boldsymbol{X}_* \boldsymbol{\beta}_*\|^2/n + o_p(1).
\end{aligned}
$$

Then, we have

$$n^{-1}I_{11} = \log\left(1 + \frac{\hat{\sigma}_j^2 - \hat{\sigma}_*^2}{\hat{\sigma}_*^2}\right) = \log\left\{1 + \frac{\|\boldsymbol{X}_* \boldsymbol{\beta}_* - \boldsymbol{H}_j \boldsymbol{X}_* \boldsymbol{\beta}_*\|^2}{n\sigma^2}\right\} + o_p(1), \qquad (21)$$

and it follows from the assumption (A3) that

$$\liminf_{n\to\infty} \log\left\{1 + \frac{\|\boldsymbol{X}_* \boldsymbol{\beta}_* - \boldsymbol{H}_j \boldsymbol{X}_* \boldsymbol{\beta}_*\|^2}{n\sigma^2}\right\} > 0. \qquad (22)$$

Because $\boldsymbol{y}^t \boldsymbol{A}_j \boldsymbol{y}/(n\hat{\sigma}_j^2) = 1 + o_p(1)$ and $\boldsymbol{y}^t \boldsymbol{A}_* \boldsymbol{y}/(n\hat{\sigma}_*^2) = 1 + o_p(1)$, we obtain

$$n^{-1}I_{12} = o_p(1). \qquad (23)$$

18

Second, we evaluate I_2. It follows that

$$\log|\boldsymbol{X}_j^t\boldsymbol{X}_j + \boldsymbol{W}_j^{-1}| = p_j \log n + \log|\boldsymbol{X}_j^t\boldsymbol{X}_j/n + \boldsymbol{W}_j^{-1}/n| = p_j \log n + O(1).$$

In addition, $\log|\boldsymbol{X}_*^t\boldsymbol{X}_* + \boldsymbol{W}_*^{-1}| = p_* \log n + O(1)$. Then,

$$n^{-1}I_2 = (p_j - p_*)n^{-1}\log n + o(1) = o(1). \tag{24}$$

Lastly, it is easy to see that

$$n^{-1}I_3 = o(1). \tag{25}$$

From (21)–(25), it follows that

$$P\{\mathrm{IC}_{\pi,1}(j) - \mathrm{IC}_{\pi,1}(j_*) > 0\} \to 1, \tag{26}$$

for all $j \in \mathcal{J}-$.

[Case of $j \in \mathcal{J}_+ \setminus \{j_*\}$]. First, we evaluate I_1. From

$$\hat{\sigma}_*^2 - \hat{\sigma}_j^2 = \boldsymbol{\varepsilon}^t(\boldsymbol{H}_j - \boldsymbol{H}_*)\boldsymbol{\varepsilon}/n = O_p(n^{-1}), \tag{27}$$

it follows that

$$(\log n)^{-1}I_{11} = (\log n)^{-1} \cdot n \log\left\{\frac{\hat{\sigma}_*^2 - (\hat{\sigma}_*^2 - \hat{\sigma}_j^2)}{\hat{\sigma}_*^2}\right\}$$
$$= (\log n)^{-1} \cdot n \cdot \log\{1 + O_p(n^{-1})\} = o_p(1). \tag{28}$$

For I_{12}, from (27) and $\boldsymbol{y}^t\boldsymbol{A}_j\boldsymbol{y} - \boldsymbol{y}^t\boldsymbol{A}_*\boldsymbol{y} = O_p(1)$, we obtain

$$I_{12} = \boldsymbol{y}^t\boldsymbol{A}_j\boldsymbol{y}/\hat{\sigma}_j^2 - \boldsymbol{y}^t\boldsymbol{A}_*\boldsymbol{y}/\hat{\sigma}_*^2$$
$$= (\boldsymbol{y}^t\boldsymbol{A}_j\boldsymbol{y} - \boldsymbol{y}^t\boldsymbol{A}_*\boldsymbol{y})/\hat{\sigma}_*^2 + O_p(1) = O_p(1).$$

Then,

$$(\log n)^{-1}I_{12} = o_p(1). \tag{29}$$

Second, we evaluate I_2. Since $p_j > p_*$ for all $j \in \mathcal{J}_+ \setminus \{j_*\}$,

$$\liminf_{n\to\infty}(\log n)^{-1}I_2 = p_j - p_* > 0. \tag{30}$$

Finally, it is easy to see that

$$(\log n)^{-1}I_3 = o(1). \tag{31}$$

From (28)–(31), it follows that

$$P\{\mathrm{IC}_{\pi,1}(j) - \mathrm{IC}_{\pi,2}(j_*) > 0\} \to 1, \tag{32}$$

for all $j \in \mathcal{J}_+ \setminus \{j_*\}$.

Combining (26) and (32), we obtain

$$P\{\mathrm{IC}_{\pi,1}(j) - \mathrm{IC}_{\pi,1}(j_*) > 0\} \to 1,$$

for all $j \in \mathcal{J} \setminus \{j_*\}$, which shows that $\mathrm{IC}_{\pi,1}$ is consistent. $\qquad\square$

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *2nd International Symposium on Information Theory*, (B.N. Petrov and F. Csaki, eds.), 267–281, Akademiai Kiado, Budapest.

Akaike, H. (1974). A new look at the statistical model identification. System identification and time-series analysis. *IEEE Transactions on Automatic Control*, **AC-19**, 716–723.

Akaike, H. (1980a). On the use of predictive likelihood of a Gaussian model. *Annals of the Institute of Statistical Mathematics*, **32**, 311–324.

Akaike, H. (1980b). Likelihood and the Bayes procedure. In *Bayesian Statistics*, (Bernard, N.J., Degroot, M.H., Lindley, D.V. and Smith, A.F.M. eds.), Valencia, Spain, University Press, 141–166.

Ando, T. (2007). Bayesian predictive information criterion for the evaluation of hierarchical Bayesian and empirical Bayes models. *Biometrika*, **94**, 443–458.

Battese, G.E., Harter, R.M. and Fuller, W.A. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, **83**, 28–36.

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, **52**, 345–370.

Dellaportas, P., Forster, J.J. and Ntzoufras, I. (1997). On Bayesian model and variable selection using MCMC. *Technical report, Department of Statistics, Athens University of Economics and Business*, Athens, Greece.

George, E.I. and McCulloch, R.E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, **88**, 881–889.

George, E.I. and McCulloch, R.E. (1997). Approaches for Bayesian variable selection. *Statistica Sinica*, **7**, 339–373.

Henderson, C.R. (1950). Estimation of genetic parameters. *The Annals of Mathematical Statistics*, **21**, 309–310.

Hurvich, C.M. and Tsai, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika*, **76**, 297–307.

Kitagawa, G. (1997). Information criteria for the predictive evaluation of Bayesian models. *Communications in Statistics — Theory and Methods*, **26**, 2223–2246.

Kuo, L. and Mallick, B. (1998). Variable selection for regression models. *Sankhya, series B*, **60**, 65–81.

Nishii, R. (1984). Asymptotic properties of criteria for selection of variables in multiple regression. *The Annals of Statistics*, **12**, 758–765.

O'Hara, R.B. and Sillanpaa, M.J. (2009). A review of Bayesian variable selection methods: what, how and which. *Bayesian Analysis*, **4**, 85–118.

Patterson, H.D. and Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika*, **58**, 545–554.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, **6**, 461–464.

Shao, J. (1997). An asymptotic theory for linear model selection. *Statistica Sinica*, **7**, 221–264.

Shi, P. and Tsai, C.-L. (2002). Regression model selection—a residual likelihood approach. *Journal of the Royal Statistical Society series B*, **64**, 237–252.

Shibata, R. (1981). An optimal selection of regression variables. *Biometrika*, **68**, 45–54.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P. and Van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society series B*, **64**, 583–639.

Srivastava, M.S. and Kubokawa, T. (2010). Conditional information criteria for selecting variables in linear mixed models. *Journal of Multivariate Analysis*, **101**, 1970–1980.

Sugiura, N. (1978). Further analysis of the data by Akaike's information criterion and the finite corrections. *Communications in Statistics — Theory and Methods*, **7**, 13–26.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society series B*, **58**, 267–288.

Vaida, F. and Blanchard, S. (2005). Conditional Akaike information for mixed-effects models. *Biometrika* **92**, 351–370.

Xu, X. and Ghosh, M. (2015). Bayesian variable selection and estimation for group lasso. *Bayesian Analysis*, **10**, 909–936.

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*, (Goel, P.K. and Zellner, A., eds.), 233–243, Amsterdam: North-Holland/Elsevier.